Alignment of CNN and Human Judgments of Geometric and Topological Concepts

Neha Upadhyay[1], **Vijay Marupudi**[1], **Kamala Varma**[2], **Sashank Varma**[1]

[1]Georgia Institute of Technology
[2]University of Maryland, College Park

Abstract

AI and ML are poised to provide new insights into mathematical cognition and development. Here, we focus on the domains of geometry and topology (GT). According to one prominent developmental perspective, infants possess core knowledge of GT concepts, presumably underwritten by dedicated neural circuitry. We use the alignment between human cognition and computer vision models to evaluate an alternate proposal: that these concepts are learned "for free" through experience with the visual world. Specifically, we measure the sensitivity of five convolutional neural network (CNN) models to 43 GT concepts that aggregate into seven classes. We focus on CNNs over other architectures (e.g., vision transformers) because their neural plausibility has been established through studies mapping their layers to areas of the brain's ventral visual stream. We find evidence that the CNNs are sensitive to some classes (e.g., Euclidean Geometry) but not others (e.g., Geometric Transformations). The models' sensitivity is generally lower at lower layers and maximal at the final fully-connected layer. Experiments with models from the ResNet family show that increasing model depth does not necessarily increase sensitivity to GT concepts. The models' profiles of sensitivity to the seven classes roughly align with the profile shown by humans, with ResNet-18 corresponding best to Western adults and DenseNet to Western children ages 3 – 6 years. This case study shows how CNNs can provide evidence for the learnability of mathematical concepts and thus inform theoretical debates in cognitive and developmental science. These findings set the stage for future experiments with other vision model architectures.

Introduction

Humans and many other animals are sensitive to geometric and topological (GT) concepts. This may be because the environment has spatial structure (Gibson 1979; Shepard 2001) and navigating this structure is important for survival (Vallortigara 2017). Thus, evolutionary processes may have ensured that understanding of GT concepts is part of infants' "core knowledge" (Spelke and Kinzler 2007). Cognitive science research has investigated understanding of GT concepts in human adults and children, non-human primates, and other animals. These studies have found evidence of sensitivity to shape (e.g., quadrilateral vs. circle), angle, convexity, rotation, and translation (Chiandetti and Vallortigara 2008; Hung et al. 2005; Pasupathy and Connor 1999). A dominant view is that many GT concepts are "intuitive" and that learning plays a minor role in human sensitivity to them (although this sensitivity may come online at different times during development). These can be seen as a modern version of the classical argument from Plato's dialogue, *Meno*, which states that humans do not "learn" GT concepts through experience and instruction, but rather already "know" them, and must merely "recollect" them (Hohol 2019).

Here, we evaluate an opposing view: that many GT concepts are learned through visual experience in the world. We do so by asking if convolutional neural network (CNN) models of computer vision learn human-like sensitivity to GT concepts. If they do, then it is more parsimonious to assume that this sensitivity is acquired through visual experience in the world rather than to posit that people possess core knowledge of these concepts (Spelke and Kinzler 2007).

CNNs are typically trained on large datasets such as ImageNet (Deng et al. 2009). Given a new image (e.g., of a robin), they can learn to classify what is depicted (e.g., as a bird or a robin). The process of training a CNN to perform image classification differs from how humans and other animals develop vision. And yet, there are surprising correspondences between the solutions offered by CNNs and the human visual system. For example, CNNs show typicality gradients similar to those of humans when categorizing exemplars, e.g., finding that robins are "better" examples of the bird category than penguins (Battleday, Peterson, and Griffiths 2020; Vemuri, Shah, and Varma 2024). To take an example from mathematical cognition, CNNs are sensitive to the numerosity (i.e., number of items) of an image. Their latent representations are consistent with the "mental number line" documented in humans (Nasr, Viswanathan, and Nieder 2019; Upadhyay and Varma 2023). At the neural level, the layers of CNNs can be mapped to areas of the ventral visual stream. Patterns of activation on these layers are correlated with activation levels in these areas as measured by fMRI when CNNs and people process the same image (Kriegeskorte 2015; Yamins and DiCarlo 2016).

Here, we explore whether the intuitiveness of GT concepts for humans is less a matter of "core knowledge" and more accurately explained as a consequence of learning

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

to classify natural images (Greenough, Black, and Wallace 1987). We consider data from behavioral experiments investigating the sensitivity of Western and non-Western adults and children to GT concepts. We use CNNs as proxies for a learned experience of the visual world and we evaluate whether they show human-like profiles of sensitivity to the same GT concepts.

Related Work

Human Intuitions About Geometric and Topological Concepts Dehaene et al. (2006) introduced an important innovation in the study of human sensitivity to GT concepts. They developed a purely visual task to test intuitions about these concepts, one that could be administered even to people who lacked formal education in geometry. In their odd-one-out task, each stimulus contains 6 images. Five embody the target GT concept, whereas the sixth does not. The goal is to identify the "odd" or "strange" one. Figure 1(a) shows the stimulus for the GT concept "symmetry – oblique". The task contains items for 43 individual concepts. These can be grouped into seven broader classes: Topology, Euclidean Geometry, Geometrical Figures, Symmetrical Figures, Chiral Figures, Metric Properties, and Geometrical Transformations. (See Dehaene et al. (2006) for a list of all concepts and the classes to which they belong.)

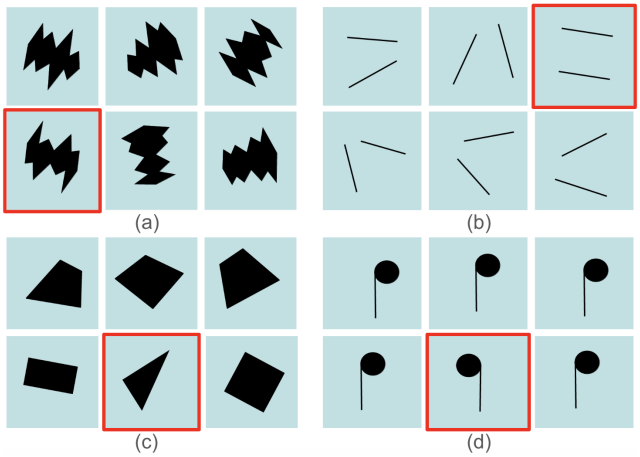

Figure 1: Dehaene et al. (2006) odd-one-out stimuli for (a) Symmetrical figures - oblique axis (b) Euclidean geometry - secant lines (c) Geometrical figures - quadrilaterals and (d) Chiral figures - vertical axis.

Dehaene et al. (2006) administered the odd-one-out task to the Mundurucu, an indigenous Amazonian group whose members live in isolated villages and do not have access to formal schooling. Both adults and children showed above-chance sensitivity to 39 of the 43 GT concepts, with the adults and children performing similarly. The researchers also administered this task to American participants. The American adults showed greater sensitivity than the American school-age children, who performed comparably to the Mundurucu adults and children. Izard and Spelke (2009) replicated these findings for American adults. They also tested a group of much younger children, ages 3 – 6 years, who showed above-chance sensitivity to fewer (i.e., 27) of the 43 concepts. The improvement in sensitivity with development may suggest a role for learning for some concepts.

An interesting finding is that, across studies, samples, and tasks, people are more sensitive to some GT concepts than others. For example, at the class level, the Mundurucu have the highest accuracy on Euclidean Geometry concepts (84%) and the lowest on Geometrical Transformations concepts (35%). Similarly, young children are above-chance on all eight of the Euclidean Geometry concepts but none of the eight Geometrical Transformations concepts (Izard and Spelke 2009). Marupudi and Varma (2023) developed a 2-Alternative Forced Choice variant of the original task where participants view a target image and must judge which of the two alternative images is more similar to it. The target image and one alternative image (the correct response) embody the GT concept whereas the other alternative image (incorrect response) does not. They found that American adults were most accurate on the Euclidean Geometry concepts and least accurate on the Geometrical Transformations concepts.

CNN Models of Computer Vision Throughout this work, we focus on CNNs that have been pre-trained on ImageNet (Deng et al. 2009), a large-scale dataset commonly used as an object classification benchmark. This task involves assigning a class label to an image, where the label corresponds to the category of an object in the image. Performing this task requires learning *visual concepts* from images. This is highly relevant to our work because object classification is a fundamental skill for both computer vision and human vision. The large size of ImageNet enables models trained on this dataset to generalize relatively well to other vision tasks.

Sensitivity of CNNs to GT Concepts Previous work has evaluated the sensitivity of CNNs to image transformations and object geometry. Mumuni and Mumuni (2021) review some of the most promising approaches to extending CNN architectures to handle non-trivial geometric transformations. However, most approaches aim to make models invariant, and therefore insensitive, to these transformations. For example, affine transformations preserve parallel lines, and therefore CNNs trained to perform such transformations are sensitive to parallel lines, but are insensitive to the affine transformations themselves (Jaderberg, Simonyan, and Zisserman 2015). To take another example, including rotations of images in the training data enables CNNs to become invariant to the 'rotation' concept (Laptev et al. 2016). Moving beyond Geometric Transformation concepts, other researchers have shown how to train CNNs to be invariant to Symmetry concepts (Cohen, Geiger, and Weiler 2019). Heinke et al. (2021) tested whether, unlike the human vision system, CNNs do not compute representations of global and local object geometry during image classification. They trained and tested six CNNs (AlexNet, VGG-11, VGG-16, ResNet-18, ResNet-50, GoogLeNet) to discriminate geometrically possible and impossible objects and compared the results with human performance. Unlike human observers,

none of the CNNs could reliably discriminate between possible and impossible objects. This finding revealed that, unlike human vision, CNNs do not compute representations of object geometry. Instead, they appear to rely strongly on local image information (Baker et al. 2020; Geirhos et al. 2019). However, other researchers have found evidence that CNNs compute global and local object geometry. Zeiler and Fergus (2014) used a DeconvNet to trace back activations in feature maps to the input image pixels that gave rise to them. They showed that the model was not just making use of broad scene context; it was also highly responsive to local geometry. They also showed that CNNs inherently process images hierarchically, with earlier layers capturing lower-level features of an image (e.g., basic shapes) and later layers capturing higher-level features. Thus, we might expect to see greater sensitivity to GT concepts at later layers, a prediction we evaluate in this work. Most similarly to the current work, Hsu, Wu, and Goodman (2022) tested the sensitivity of CNN models to Euclidean geometry concepts. Models and participants were provided 5 target images that embodied the concepts and were required to identify concept matches from a set of 15 images which included 5 concept matches and 10 distractors that were either close to or far from the target concept. While adults completed the task with high accuracy, CNNs pre-trained on ImageNet performed poorly. More recently, Campbell et al. (2024) found that a vision transformer model (DINOv2) and a multi-modal model (CLIP) showed human-like sensitivity to some Geometrical Figures and concepts (i.e., 'square', 'rectangle', 'parallelogram'), whereas a CNN (ResNet-50) showed reduced sensitivity comparable to that of baboons.

In this study, we seek to compare the performance of these models to a wider array of Geometric and Topological concepts using standard items used in prior studies (Dehaene et al. 2006), allowing us to compare the performance of CNN models to more diverse human groups.

Research Questions

The current study investigated the sensitivity of CNN models to GT concepts. It addressed five research questions:

1. Are CNNs sensitive to the 43 GT concepts, aggregated into 7 classes, that have been evaluated in humans?
2. Does the GT-sensitivity of CNNs vary across model layers?
3. Are the accuracy profiles of CNNs across the 7 classes correlated with those of adults and/or children?
4. Does the GT-sensitivity of CNNs increase for models of greater depth?
5. Do different CNNs show similar or different profiles of sensitivity and insensitivity across the 7 classes?

Method

Models

We chose five well-known CNNs for our experiments: **AlexNet** (Krizhevsky, Sutskever, and Hinton 2012), **VGG-19** (Simonyan and Zisserman 2015), **GoogLeNet** (Szegedy et al. 2015), **ResNet-18** (He et al. 2016), and, **DenseNet-121**(Huang et al. 2017). We accessed versions of these models that have been pre-trained on ImageNet (Deng et al. 2009) through the Keras API (Chollet et al. 2015).

The use of models pre-trained on ImageNet offers a pertinent comparison to the Mundurucu adults and children and to the young Western children. Both groups exhibit expert-level performance in object classification, and yet, both have minimal exposure to the abstract geometric and topological stimuli used in this study. Thus, their performance may be equitably compared to that of CNNs.

Stimuli and Datasets

The Dehaene et al. (2006) odd-one-out task includes one stimulus for each of the 43 concepts. Each stimulus consists of 6 images: 5 that embody the concept and 1 that does not. The 43 concepts aggregate into 7 broader classes: Topology, Euclidean Geometry, Geometrical Figures, Symmetrical Figures, Chiral Figures, Metric Properties, and Geometrical Transformations.

The human data for the odd-one-out task comes from two prior cognitive science studies:

- Dehaene et al. (2006) reported the performance of adults and children belonging to the Mundurucu, an indigenous Amazonian group. The participants in this case were inhabitants of isolated villages and had received little to no formal education.
- Izard and Spelke (2009), in their first experiment, reported the performance of Western children between the ages of 3 – 6 years. In their second experiment, they reported the performance of 400 Western children and adults ages 6 – 51 years. We requested the data for adults ages 18 – 25 years.

These datasets represent profiles of sensitivity and insensitivity to the 43 GT concepts and the 7 classes across different cultural groups (Mundurucu, Western) and different age groups (adults, children).

Marupudi and Varma (2023) showed that these profiles correlate across datasets at approximately the $r = 0.7$ level. However, there are some interesting differences, particularly between the profiles of adults and children, that we consider below.

Deriving Model Predictions

Each stimulus consists of 6 images; see Figure 1(a). Each image was re-scaled and cropped to a size of 224×224. We presented each image to a given model and recorded the vector of activations obtained from every layer. As a measure of a model's "overall" performance, we considered the predictions made by the final fully-connected layer.

We explored various methods for computing the odd-one-out among the 6 images of each stimulus, including applying outlier detection algorithms such as Local Outlier Factor and also K-Means clustering. The small number of images in each stimulus made such approaches unstable. We ultimately settled on the following simple procedure for identifying the image that was *least similar* to the others: For each image, we computed the cosine similarity between its

Model	Topology	Euclidean Geometry	Geometrical Figures	Symmetrical Figures	Chiral Figures	Metric Properties	Geometrical Transformations	Overall Accuracy	Avg. Rank
VGG-19	50.0	37.5	22.22	0.0	75.0	14.29	12.5	31.11*	3.98
AlexNet	25.0	50.0*	33.33	33.33	50.0	14.29	12.5	33.33*	3.82
ResNet-18	**25.0**	**75.0***	**33.33**	**33.33**	**50.0**	**28.57**	**12.5**	**40***	**4.27**
DenseNet	25.0	50.0*	33.33	0.0	0.0	28.57	12.5	28.89*	4.07
GoogLeNet	25.0	62.5*	33.33	0.0	50.0	42.86	12.5	37.78*	4.18

Table 1: Overall performance of the CNN models on the odd-one-out task. Entries that differ significantly from the chance level of 16.67% (i.e., where $p < 0.05$ on a binomial test) have been marked with an "*". The results of the overall best performing model, ResNet-18, are shown in bold.

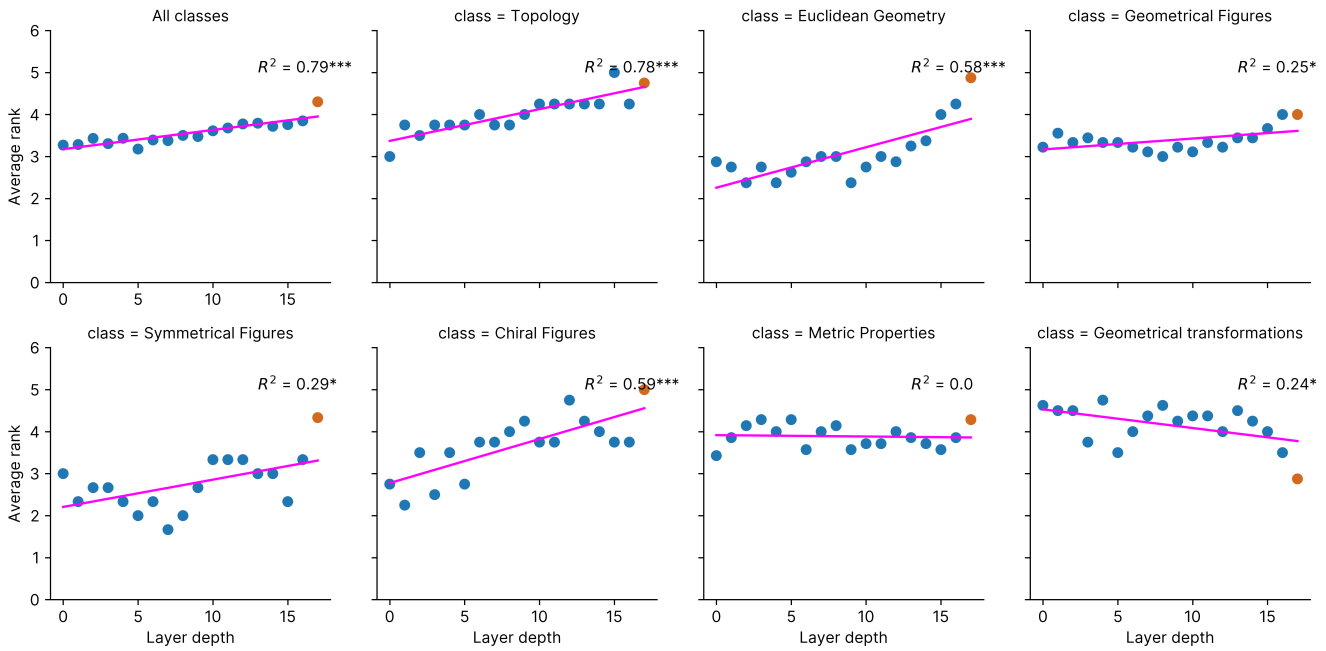

Figure 2: Average rank across layers of ResNet-18 (higher is better). The final fully-connected layer is colored orange.

vector representation and the vector representations of each of the other 5 images. We then computed its average cosine similarity to the other 5 images. (We note that this simple procedure is the same as that adopted by Campbell et al. (2024), Muttenthaler et al. (2023a), and Muttenthaler et al. (2023b)). Finally, we rank-ordered the six images from the one with the highest average similarity to all other images (rank = 1) to the one with the lowest average similarity (rank = 6).

We quantified model performance in two ways. The first was accuracy: whether the odd-one-out image was correctly identified (i.e., had rank = 6) or not. This is a binary outcome. In what follow, we report the average accuracy computed overall, across all 43 concepts, or computed for each of the 7 classes. We also computed the average rank of the odd-one-out image, with higher values indicating better model performance (i.e., assignments closer to rank = 6).

We evaluated model performance using different correlation statistics, i.e., Pearson and point-biserial.

Results

Overall Performance of the Models

Table 1 summarizes the overall performance of the CNNs (i.e., as measured on the final fully-connected layer) on the odd-one-out task, aggregated to the level of the 7 classes of GT concepts. There are three important patterns to notice. First, the best-performing model is ResNet-18: it has the highest overall accuracy (40%). (Note that it also assigns the odd-one-out image the highest average rank (4.27; correct performance is 6).) Second, even ResNet-18 performs much worse than humans. Its accuracy (40%) is far below that achieved by the Mundurucu adults and children of the original Dehaene et al. (2006) study (90.7%) and by the young Western children ages 3 – 6 years of the Izard and Spelke (2009) study (62.8%). Third, the models differ in their average accuracy across the 7 classes. For example, ResNet-18 is relatively accurate on Euclidean Geometry concepts (75%) but relatively inaccurate on Geometrical Transforma-

tions concepts (12.5%). Thus, there is mixed evidence that the CNNs can match the high levels of sensitivity to GT concepts shown by humans purely through training on naturalistic images.

To further probe the correspondence between their performance profiles and those of humans, the next set of experiments focused on ResNet-18, which offered the best overall performance.

Sensitivity of ResNet-18 to GT Concept Classes

The above analyses reported ResNet-18's performance on its final fully-connected layer. We expected representations at this level to show the greatest sensitivity to GT concepts. We also considered the performance of each of its layers on each of the 7 classes. Our prediction was that performance would be relatively poor for the earliest, convolutional layers and would increase as the model further processed the images. For these analyses, we used the finer-grained performance measure of average rank of the odd-one-out image. Recall that correct identification of this image is signified by $rank = 6$. The results are shown in Figure 2.

As expected, the predicted rank of the odd image generally increases (i.e., ResNet-18 becomes more accurate) across the layers. This positive linear association held for 5 of the 7 classes of GT concepts. There was no such association for the Metric Properties class. Note that this was *not* because ResNet-18 had low overall accuracy on this class. In fact, it had lower overall accuracy on the Topology and Geometric Transformations classes; see Table 1. For the Geometric Transformations class, the association was negative, with more accurate model predictions at the lower layers. This is likely due to the practice, when training models to perform image classification, of augmenting the training data with some of the transformations tested by the Geometrical transformations class, to induce invariance.

To complement this analysis of ResNet-18's performance by predicted rank, we also evaluated its absolute accuracy in correctly identifying the odd image, i.e., assigning it $rank = 6$. Figure 3 shows the accuracy of the model across all of its layers, separately for each of the 7 classes and then averaged across all stimuli. Again, our prediction was that accuracy would increase from earlier layers to later layers and would be maximal at the final fully-connected layer. There was some evidence for this prediction for this more stringent performance measure of absolute (vs. rank) accuracy.

Correspondence Between the CNN and Human Profiles

We next evaluated whether ResNet-18's accuracy profile across the 7 classes corresponded to the profiles shown by human participants in three of the behavioral samples reported above: the Izard and Spelke (2009) experiment with Western children ages 3 – 6 years and Western adults ages 18-25 years, and the Dehaene et al. (2006) study of Mundurucu adults and children. All of these studies used the odd-one-out task. Figure 4 shows the class-level accuracy profile for ResNet-18 computed on the final fully-connected layer. It also shows the profiles for the three behavioral samples.

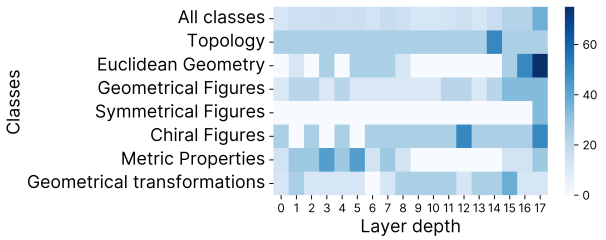

Figure 3: Percentage of correct predictions across layers of ResNet-18.

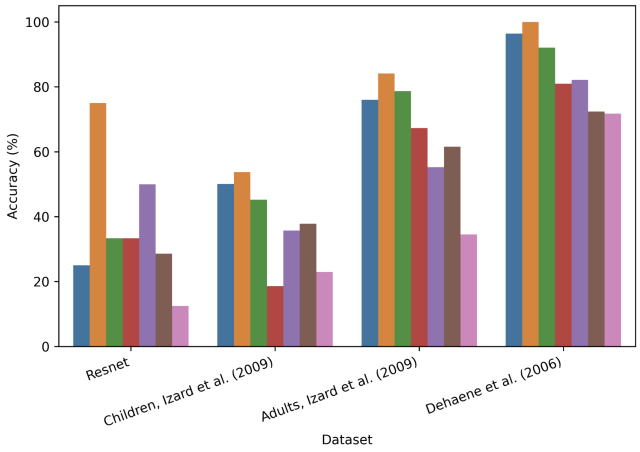

Figure 4: Absolute accuracies across the 7 classes for ResNet-18 and for 3 behavioral samples on the odd-one-out task. Pearson correlations are summarized in Table 2.

Although the model's accuracies are generally lower than those of humans, the profiles of the models and the humans across the 7 classes are comparable. For example, both ResNet-18 and humans tend to perform best on the Euclidean Geometry class and worst on the Geometrical Transformations class. Table 2 quantifies the correspondence between the model and human profiles. ResNet-18 correlates $r = .52$ with children ages 3 – 6 years old from Izard and Spelke (2009), $r = .60$ with adults ages 18-25 years from the same study, and $r = .58$ with Mundurucu children and adults from Dehaene et al. (2006). Notably, two other models, DenseNet and GoogLeNet, achieve high correlations with children ages 3 – 6 years old ($r = 0.81$ and $r = 0.74$, respectively).

Table 2 also shows the correspondence between all 5 models and the 3 human profiles. Each cell contains the Pearson correlation between the accuracy profiles across the 7 classes for a model's final fully-connected layer and a dataset. The correlations are highest for DenseNet, indicating alignment between the model and humans about the relative difficulty of different classes. The alignment is par-

Model	Children, Izard and Spelke (2009)	Adults, Izard and Spelke (2009)	Dehaene et al. (2006)
VGG19	0.51	0.41	0.13
ResNet-18	**0.52**	**0.6**	**0.58**
AlexNet	0.36	0.61	0.50
DenseNet	0.81	0.59	0.62
GoogLeNet	0.74	0.40	0.38

Table 2: Pearson correlations between the final fully-connected layers of the 5 models and humans from 3 behavioral samples on the odd-one-out task, computed across the 7 classes.

ticularly close between DenseNet and the Izard and Spelke (2009) data collected from children ages 3 – 6 years.

Model Depth

We next investigated whether the correspondence between model and human profiles increases for models of greater depth. We again focused on the ResNet family of models. To ResNet-18, the model considered above, we added ResNet-34, ResNet-50, ResNet-101 and ResNet-152. Their performance on the odd-one-out task was computed at their final fully-connected layer and aggregated to the level of the 7 classes. The results are shown in Table 3.

There are two patterns to notice. First, the shallowest model of this family, ResNet-18, offered arguably the best performance. It achieved the highest overall accuracy, correctly identifying the odd image for 40% of the GT concepts. Second, the deepest model. ResNet-152, offered the second-best performance. This was driven by its relative success on the Geometrical Figures class (55.56% vs. 33.33% for ResNet-18) and the Geometrical Transformations class (25% vs. 12.5% for ResNet-18). It also assigned the odd image the highest rank on average, signifying global sensitivity to GT concepts. (Recall that correct identification occurs when the odd image is assigned $rank = 6$.)

This experiment shows that although increasing depth might improve performance on object classification, it doesn't necessarily improve human alignment. Instead, alignment may be more dependent on the model's objective function rather than its architecture or number of layers or parameters (Muttenthaler et al. 2022).

It is interesting to note that of the ResNet variants, only ResNet-18 scored well on Symmetry concepts. It is possible that the small size and architecture of ResNet-18 forced it to extract symmetry as a latent concept. By contrast, the larger variants may have been able to achieve good classification performance via more 'brute force' strategies. This suggests that network size might be acting as a useful constraint that results in the development of more abstract representations of GT concepts, an observation that has been made by cognitive scientists modeling language development (Elman 1993).

Model Comparison

Although the 5 models are all CNNs, there are important differences in their architectures. This raises the question of the degree to which their predictions agree or disagree with each other. We evaluated this by comparing the performance profiles of each pair of models, as derived from their final fully-connected layers, across the 7 classes. The Pearson correlations are collected in Table 4. The most prominent pattern is the high correlations among ResNet-18, AlexNet, and GoogLeNet, i.e., strong agreement on the relative difficulty of different classes of GT concepts.

Discussion

This study investigated the sensitivity of CNNs to geometric and topological concepts. This class of vision models is of particular interest to cognitive science because of their potential neural plausibility (Kriegeskorte 2015; Yamins and DiCarlo 2016). Specifically, it addressed five research questions.

First, it showed that CNNs are sensitive to the same GT concepts as humans. Every model achieved above-chance performance on the odd-one-out task (greater thank 16.67%; Table 1), with ResNet-18 achieving the highest overall accuracy (40%). Importantly, the classes ResNet-18 found easiest (Euclidean Geometry) and most difficult (Geometric Transformations) are also the ones young children find easiest and most difficult (Izard and Spelke 2009). That said, all models showed a notable lack of sensitivity to concepts from the Symmetrical Figures and Geometric Transformations classes.

Second, focusing on the ResNet-18 model that achieved the highest overall accuracy, it showed that GT sensitivity differs across layers. As predicted, accuracy was higher for later layers and highest for the final fully-connected layer (see Figures 2 and 3). This was true for all but 2 of the 7 classes (Metric Properties and Geometric Transformations).

Third, it showed that the accuracy profiles of the models across the 7 classes, as derived from their final fully-connected layers, correlated with those of humans (Figure 4 and Table 2). ResNet-18 achieved moderate correlations with the profiles of both adults and children ages 3 – 6 years. Also notable was that the DenseNet and GoogLeNet models achieved high correlations with the profiles of the children.

Fourth, it showed that the performance of the models does *not* improve with increasing depth. Among the ResNet family of models, the shallowest model (ResNet-18) generally outperformed the deeper models (Table 3).

Fifth, it showed that the models, which are all CNNs, generally agree with each other on which of the 7 classes are harder and which are easier (Table 4). This was most strongly true for ResNet-18, AlexNet, and GoogLeNet.

Implications for Cognitive Science

These findings show that CNNs have similar sensitivity profiles to classes of GT concepts as humans. This suggests that learning to categorize the world requires a certain amount of sensitivity to GT concepts and that this sensitivity can be learned. This stands as an alternative to proposals in cognitive science, developmental science, and neuroscience that GT concepts are 'intuitive' and are known independent of visual experience. In particular, it contrasts with the proposal that infants possess "core knowledge" of GT concepts

Model	Topology	Euclidean Geometry	Geometrical Figures	Symmetrical Figures	Chiral Figures	Metric Properties	Geometrical Transformations	Overall Accuracy	Avg. Rank
ResNet-18	**25.0**	**75.0**	**33.33**	**33.33**	**50.0**	**28.57**	**12.5**	**40**	**4.27**
ResNet-34	0.0	50.0	22.22	0.0	25.0	28.57	0.0	24.44	4.18
ResNet-50	25.0	50.0	33.33	0.0	25.0	42.86	12.5	33.33	4.18
ResNet-101	50.0	37.5	44.44	0.0	50.0	14.29	12.5	33.33	4.24
ResNet-152	25.0	50.0	55.56	0.0	25.0	28.57	25.0	37.78	4.47

Table 3: Overall performance of ResNet models of increasing depth on the odd-one-out task.

Model	VGG19	ResNet-18	AlexNet	DenseNet	GoogLeNet
VGG19	1.00	0.42	0.58	-0.03	0.60
ResNet-18	0.42	1.00	0.88	0.43	0.75
AlexNet	0.58	0.88	1.00	0.08	0.54
DenseNet	-0.03	0.43	0.08	1.00	0.59
GoogLeNet	0.60	0.75	0.54	0.59	1.00

Table 4: Correlations between the predicted accuracies across the 7 classes derived from the final fully-connected layer for the 5 models.

to bootstrap their development (Spelke and Kinzler 2007). Our experiments provide evidence that such concepts can be learned through visual experience without having to posit a representational bias. That said, "core knowledge" may still be important for improving data efficiency, enabling children to learn more from fewer observations.

One direction for future research is modeling the development of GT concepts. The current experiments show that CNNs trained on the classification of naturalistic images learn some GT concepts. However, the overall accuracy of ResNet-18, the best performing model, was still below that of children ages 3 – 6 years old, 40% vs. 62.7% (Izard and Spelke 2009). Future work could begin with ResNet-18 and investigate whether, through fine-tuning on more advanced visual tasks, the model develops stronger sensitivities to GT concepts and along roughly the same developmental progression as humans. Another direction for future research is to use CNNs to make neural predictions. Vision scientists have found promising correspondences between (i) the activations of the layers of CNNs while processing images and (ii) neural activity in the corresponding brain areas of the ventral visual stream while viewing the same images in the MRI scanner (Kriegeskorte 2015; Yamins and DiCarlo 2016). Thus, the current results can be used to make predictions about which brain areas are sensitive to which classes of GT concepts.

Implications for Machine Learning

Human-level performance is often a standard for evaluating the performance of ML models. The current research encourages the use of cognitive benchmarks to evaluate whether the latent representations of computer vision models are human-like. Our findings provide insights into what CNNs learn, which can be useful in a variety of ways.

First, they can potentially guide model selection. For example, if a CNN needs to learn an image classification task where the orientation of an image is a key distinguishing feature, then it would be advantageous to select a model architecture that is more sensitive to Geometrical Transformations. More generally, our findings provide an additional perspective on why some architectures perform better on some tasks than others.

Second, our study exposes model weaknesses that can be important directions for further exploration. One example is the blind spot that all models exhibited for Symmetry concepts. This paves the way for what we might expect by choosing different model architectures. For example, Vision Transformers differ from CNNs in that they are not inherently translation-invariant and do not have an inductive locality bias – two qualities that likely influence their sensitivity to GT concepts. Whether this is the case, and they show increased sensitivity to GT concepts, is a question that demands further research. In ongoing work, we are evaluating the potential of vision transformers for understanding human sensitivity to GT concepts. The early results are promising. For example, the ViT (Dosovitskiy et al. 2021) and DINOv2 (Oquab et al. 2024) models achieve higher overall accuracies on the 43 GT concepts than ResNet-18 (48% and 50%, respectively, vs. 40%).

Third, in the current study we did not fine-tune the ImageNet-trained CNNs. This choice was dictated by our scientific goal of evaluating whether models incidentally learn the GT concepts relevant to their training task. This enables us to compare their learned sensitivities to human sensitivities to GT concepts in the absence of formal education (i.e., training) on them (Dehaene et al. 2006). Thus, the test images and stimuli were out-of-distribution for the CNN models, which partially explains their relatively poor performance. However, the same would be true for members of the Mundurucu and for Western children before they begin formal schooling: they have had little to no exposure to such stimuli during their lives and yet perform better than CNN models for many classes.

This raises the question of why these models perform relatively poorly. One explanation is that their training mechanisms aim to increase their invariance to certain transformations, such as rotation and translation, and this unintentionally makes them insensitive to these transformations when they are relevant for the task at hand. Future research could explore the GT sensitivities of models whose performance has been fine-tuned for the odd-one-out task. This would reveal whether further GT concepts are *latent* in the representations learned by CNNs during image classification, enabling these models to be sensitive to task-specific optimizations.

References

Baker, N.; Lu, H.; Erlikhman, G.; and Kellman, P. J. 2020. Local features and global shape information in object classification by deep convolutional neural networks. *Vision research*, 172: 46–61.

Battleday, R. M.; Peterson, J. C.; and Griffiths, T. L. 2020. Capturing Human Categorization of Natural Images by Combining Deep Networks and Cognitive Models. *Nature Communications*, 11(1): 5418.

Campbell, D.; Kumar, S.; Giallanza, T.; Griffiths, T. L.; and Cohen, J. D. 2024. Human-Like Geometric Abstraction in Large Pre-trained Neural Networks. In *46th Annual Meeting of the Cognitive Science Society*. Rotterdam, Netherlands.

Chiandetti, C.; and Vallortigara, G. 2008. Is There an Innate Geometric Module? Effects of Experience with Angular Geometric Cues on Spatial Re-Orientation Based on the Shape of the Environment. *Animal Cognition*, 11(1): 139–146.

Chollet, F.; et al. 2015. Keras.

Cohen, T. S.; Geiger, M.; and Weiler, M. 2019. A General Theory of Equivariant Cnns on Homogeneous Spaces. *Advances in neural information processing systems*, 32.

Dehaene, S.; Izard, V.; Pica, P.; and Spelke, E. S. 2006. Core Knowledge of Geometry in an Amazonian Indigene Group.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Elman, J. L. 1993. Learning and Development in Neural Networks: The Importance of Starting Small. *Cognition*, 48(1): 71–99.

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*.

Gibson, J. J. 1979. *The Ecological Approach to Visual Perception*. The Ecological Approach to Visual Perception. Boston, MA, US: Houghton, Mifflin and Company. ISBN 978-0-395-27049-3.

Greenough, W. T.; Black, J. E.; and Wallace, C. S. 1987. *Experience and Brain Development*. Brain Development and Cognition: A Reader, 2nd Ed. Malden: Blackwell Publishing. ISBN 978-0-631-21736-7 978-0-631-21737-4.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Heinke, D.; Wachman, P.; van Zoest, W.; and Leek, E. C. 2021. A failure to learn object shape geometry: Implications for convolutional neural networks as plausible models of biological vision. *Vision Research*, 189: 81–92.

Hohol, M. 2019. *Foundations of Geometric Cognition*. Routledge. ISBN 0-429-05629-X.

Hsu, J.; Wu, J.; and Goodman, N. 2022. Geoclidean: Few-Shot Generalization in Euclidean Geometry. *Advances in Neural Information Processing Systems*, 35: 39007–39019.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708.

Hung, C. P.; Kreiman, G.; Poggio, T.; and DiCarlo, J. J. 2005. Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. *Science*, 310(5749): 863–866.

Izard, V.; and Spelke, E. S. 2009. Development of Sensitivity to Geometry in Visual Forms. *Human evolution*, 23(3): 213–248.

Jaderberg, M.; Simonyan, K.; and Zisserman, A. 2015. Spatial Transformer Networks. *Advances in neural information processing systems*, 28.

Kriegeskorte, N. 2015. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, 1(1): 417–446.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Laptev, D.; Savinov, N.; Buhmann, J. M.; and Pollefeys, M. 2016. Ti-Pooling: Transformation-Invariant Pooling for Feature Learning in Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 289–297.

Marupudi, V.; and Varma, S. 2023. Graded Human Sensitivity to Geometric and Topological Concepts. *Cognition*, 232: 105331.

Mumuni, A.; and Mumuni, F. 2021. CNN architectures for geometric transformation-invariant feature representation in computer vision: a review. *SN Computer Science*, 2(5): 1–23.

Muttenthaler, L.; Dippel, J.; Linhardt, L.; Vandermeulen, R. A.; and Kornblith, S. 2023a. Human alignment of neural network representations. In *11th International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, Mai 01-05, 2023*. OpenReview.net.

Muttenthaler, L.; Linhardt, L.; Dippel, J.; Vandermeulen, R. A.; Hermann, K.; Lampinen, A.; and Kornblith, S. 2023b. Improving neural network representations using human similarity judgments. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 50978–51007. Curran Associates, Inc.

Muttenthaler, L.; Linhardt, L.; Dippel, J.; Vandermeulen, R. A.; and Kornblith, S. 2022. Human alignment of neural network representations. In *SVRHM 2022 Workshop @ NeurIPS*.

Nasr, K.; Viswanathan, P.; and Nieder, A. 2019. Number Detectors Spontaneously Emerge in a Deep Neural Network

Designed for Visual Object Recognition. *Science Advances*, 5(5): eaav7903.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; HAZIZA, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*. Featured Certification.

Pasupathy, A.; and Connor, C. E. 1999. Responses to Contour Features in Macaque Area V4. *Journal of neurophysiology*, 82(5): 2490–2502.

Shepard, R. N. 2001. Perceptual-Cognitive Universals as Reflections of the World. *Behavioral and brain sciences*, 24(4): 581–601.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

Spelke, E. S.; and Kinzler, K. D. 2007. Core Knowledge. *Developmental Science*, 10(1): 89–96.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going Deeper with Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.

Upadhyay, N.; and Varma, S. 2023. CNN Models' Sensitivity to Numerosity Concepts. In *The 3rd Workshop on Mathematical Reasoning and AI at NeurIPS'23*.

Vallortigara, G. 2017. Comparative Cognition of Number and Space: The Case of Geometry and of the Mental Number Line. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 373(1740).

Vemuri, S. K.; Shah, R. S.; and Varma, S. 2024. How Well Do Deep Learning Models Capture Human Concepts? The Case of the Typicality Effect. In *46th Annual Meeting of the Cognitive Science Society*. Rotterdam, Netherlands.

Yamins, D. L. K.; and DiCarlo, J. J. 2016. Using Goal-Driven Deep Learning Models to Understand Sensory Cortex. *Nature Neuroscience*, 19(3): 356–365.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer.